Verschillende eindtoetsen, verschillende resultaten in VO?

Maastricht University ROA

Does lowering the bar help? Results from a natural experiment in high-stakes testing in Dutch primary education

Madelon Jacobs
Rolf van der Velden
Lynn van Vugt

ROA Research Memorandum

ROA-RM-2021/4

Researchcentrum voor Onderwijs en Arbeidsmarkt | ROA
Research Centre for Education and the Labour Market | ROA

Does lowering the bar help? Results from a natural experiment in high-stakes testing in Dutch primary education

Madelon Jacobs
Rolf van der Velden
Lynn van Vugt

ROA-RM-2021/4
June 2021

Research Centre for Education and the Labour Market
Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands
T +31 43 3883647 F +31 43 3884914

secretary-roa-sbe@maastrichtuniversity.nl
www.roa.nl

ISSN: 2666-8823

Abstract

Does lowering the bar help? Results from a natural experiment in high-stakes testing in Dutch primary education*

In many countries, high-stakes tests play an important role in the allocation of pupils to prestigious tracks or schools in secondary education or students to prestigious programs or colleges in tertiary education. It is not clear what would happen if the standards for these tests were systematically raised or lowered. Would that affect the subsequent educational career? This paper exploits a unique natural experiment in the Netherlands using the market entrance of two new suppliers of high-stakes tests in primary education. In the first year of introduction, these new tests were not yet properly calibrated: For one test the standards were too low, while for the other test they were too high, compared to the standards of the traditional test that continued to be the main supplier. We use high-quality register data and a within-schools-across-cohorts design to model the short- and long-term outcomes (i.e., change in teacher advice and actual track three years later) for the students that were affected by the new tests. We find evidence for short-term effects, but no evidence for long-term effects. This implies that the Dutch educational system is sufficiently flexible to allocate pupils to the appropriate track, even if a high-stakes test advice does not recommend the right track. At the same time, it also implies that lowering the bar is not a simple way to increase the share of students going to prestigious tracks.

JEL classification: I26, J24
Keywords: high-stakes testing; transition primary to secondary education; raising or lowering standards; Netherlands cohort study on education

Madelon Jacobs
Maastricht University
ROA
P.O. Box 616
NL-6200 MD Maastricht
The Netherlands
mce.jacobs@maastrichtuniversity.nl

Rolf van der Velden
Maastricht University
ROA
P.O. Box 616
NL-6200 MD Maastricht
The Netherlands
r.vandervelden@maastrichtuniversity.nl

Lynn van Vugt
Maastricht University
ROA
P.O. Box 616
NL-6200 MD Maastricht
The Netherlands
l.vanvugt@maastrichtuniversity.nl

* The research presented in this article was funded by the Netherlands Initiative for Education Research (NRO), project number 405-17-305. The authors like to thank Cees Glas, Tim Huijts, Dinand Webbink and anonymous reviewers for their helpful comments.

1. Introduction

In many countries, high-stakes tests play an important role in the allocation of pupils to prestigious tracks or schools in secondary education or students to prestigious programs or colleges in tertiary education. It is not clear what would happen if the standards for these tests were systematically raised for some students and lowered for others. Will the former group of students experience a long-term penalty in their study career and will the latter group profit? No ethical committee would allow such an experiment to take place because of the potential strong adverse effects for those students who were denied allocation to a track that would normally fit their potential.

In this paper, we use a unique natural experiment that sheds light on the effects of raising or lowering the standards of a high-stakes test. In the Netherlands, the allocation to tracks in secondary education takes place at age 12 (grade 6) and is based on the primary school teacher's advice and the results of a nationwide high-stakes test. Before the test is taken, the primary school teachers give a so-called *initial advice* based on previous performance in school. After the high-stakes test results are available, teachers have the option to adjust their initial advice and produce a *final advice*. However, the teachers are only allowed to upgrade this advice, not downgrade it.

Until 2014, the test was not mandatory, although some 90% of the schools made use of the so-called Cito-test (as of 2014 renamed into CET-test). In 2014, the Ministry of Education, Culture and Sciences passed a new law, making the test mandatory but allowing new suppliers of high-stakes tests to enter the market. Two new suppliers entered the market at that time: Route 8 and IEP. In the first year of their introduction (2014/2015) only 4% of the schools in our dataset switched to one of these new tests, but in the second year (2015/2016)

this increased to 23%. All three tests convert the overall score on the test into a so-called track recommendation: each range of the score corresponds to a certain track in secondary education. We will refer to these converted scores as the *test advice*.

In the first years of introduction, the cut-off points for these track recommendations were not yet properly calibrated for the new tests, since they had been tested on a small sample of pupils in a low-stakes setting. This implied that the cut-off points for the different track recommendations were too high in one case and too low in the other case. As teachers are only allowed to adjust their advice in an upward direction, the test advices that were systematically too high, might result in some pupils getting a final teacher's advice that is higher than expected based on their 'true' performance. After 2015/2016 this problem was solved as the test suppliers could adjust their cut-off points on the results of the high-stakes setting in the previous year. Of course, this problem did not hold for the traditional test (Cito), as this test was already calibrated in a high-stakes setting.

It is important to highlight that schools had no prior knowledge about the bias in the standards of the new tests. They acted under the assumption that the cut-off points for the track recommendations in the test were like the ones used in the traditional test.[1] As such, the change to a new test supplier in 2015/2016 is a natural experiment, where pupils in some schools received a systematically higher track recommendation compared to what they would have received had they taken the traditional test, while pupils in other schools received systematically lower track recommendations. In this paper, we compare schools that used the traditional test in 2014/2015 and switched to one of the new tests in 2015/2016 with schools that continued using the traditional test. We use a multilevel design to model the change

1. This is also illustrated by the fact that some schools switched to a stricter test and others to a more lenient one.

within schools across the two cohorts, to take account of any effect resulting from specific schools switching to one of the new suppliers.

The research questions we aim to answer are the following: *1) To what extent does the test advice vary across different types of high-stakes tests for pupils with the same performance level? 2) To what extent do these differences affect the final teacher's advice at age 12 (short-term effects)? 3) To what extent do these differences affect pupils' educational position after three years at age 15 (long-term effects)? 4) To what extent does this differ between pupils of different socio-economic background?*

We use high-quality register data from the Netherlands Cohort Study on Education (*Nationaal Cohortonderzoek Onderwijs*: NCO; for more information see Haelermans et al., 2020). This dataset enables us to track *all* pupils in the Dutch education system and assess information about the test results from the different high-stakes test suppliers in the Netherlands. This is a huge advantage over survey data, that might suffer from selection bias and lack of statistical power.

The paper is organized as follows. After the theoretical framework (Section 2), we describe the Dutch education system and the different tests that were used (Section 3). Section 4 describes the data and methods and in Section 5 we present the results. Section 6 concludes.

2. Theoretical Framework

In this paper we highlight the role of high-stakes testing in the context of the transition from primary to secondary education in an early stratifying system: The Netherlands. In the Dutch education system, this transition is based on the advice of the primary school teacher and the track recommendation resulting from a national test. As indicated above, the track

recommendation from the test (the test advice) can be used by primary school teachers to adjust their initial advice, but only in an upward direction: If the test advice is higher than the initial advice, the primary school teacher can decide to give a higher final advice, and this might in turn result in a higher track placement.

Let us first concentrate on the first step in this process: The adjustment of the final teacher's advice. This adjustment will not automatically occur for all pupils. Teachers might be more willing to revise their initial judgement for pupils from higher social strata that are considered to have a more stimulating home environment (Timmermans, De Boer, Amsing, & Van der Werf, 2018). Moreover, as both the initial teacher advice and the test advice are communicated with the pupils and their parents, this will prompt some process of negotiation. If parents are informed that the test advice is higher than the initial teacher advice, they might put pressure in adjusting the advice in an upward direction. As Boudon (1974) made clear in his classical model on the primary and secondary effects of social stratification, this pressure will be higher from parents of the higher social strata. This is caused by differences in the cost and benefit analysis that individuals from different social strata make in the educational decision process (Breen & Goldthorpe, 1997). Parents from higher social strata perceive the benefits of following a higher education track as more beneficial and the associated costs as lower than parents from lower social strata (Boudon, 1974). Moreover, according to the relative risk aversion mechanism, parents from higher strata try to avoid downward mobility for their offspring (Breen & Goldthorpe, 1997). Breen, Van de Werfhorst and Jæger (2014) extend this rational action theory by explicitly including risk aversion and time discount preferences that differ between social strata. For pupils from high-SES families, we therefore expect that the initial teacher advice will more often be adjusted, since the parents will generally insist on placement in the higher academic tracks (Dumont, Klinge, & Maaz, 2019).

For low-SES pupils we might expect that the adjustment in the teacher's advice will be less salient, since their parents are less informed about the possibilities of the educational system and therefore less likely to go against the initial teachers' advice (Forster & Van de Werfhorst, 2020). This leads to the following hypotheses:

H1: Pupils who took the more lenient test, will have their initial teacher advice more often adjusted than pupils who took the traditional test or the stricter test.

H2: This will hold more strongly for pupils from high-SES families.

The first implication of a higher primary school teacher advice might be that these pupils will initially be allocated to a higher track in secondary education. This might affect pupils' motivation and learning habits and the expectations of secondary school teachers and parents. According to Vygotsky (1978), education should provide challenges that fit pupils' zone of proximal development, to improve their individual capabilities. This relates to the 'eighty five percent rule for optimal learning' of Wilson, Shenhav, Straccia and Cohen (2019: 1) stating that "*in many situations we find that there is a sweet spot in which training is neither too easy nor too hard, and where learning progresses most quickly*". Their theory states that around eighty-five percent of the challenges should be accurate to lead to the optimal learning curve. In the case of placing pupils in tracks that are too easy, this is easy to grasp. This form of misallocation might lead to suboptimal learning and result in boredom and counter-productive learning habits (Vaisey, 2006), although opposite findings have been found as well (Elsner & Isphording, 2017).

For our paper it is more important to look at what happens when pupils are allocated into tracks above their ability. Here the theory is not conclusive. One might expect a positive effect as it provides pupils with an opportunity to develop their talents. The so-called educational self-fulfilling prophecy (Taylor, 1979) states that reinforcing pupil's belief that

they can perform at a certain level will increase their performance. Nevertheless, one can also expect negative effects when pupils become overwhelmed with the performance requirements of the track and potentially must repeat classes, switch to lower tracks or even drop out (Hardy, 2003). We assume that low-SES pupils might suffer more from the negative consequences, while high-SES pupils might profit from the positive effects. One of the reasons why this last group does not suffer from the negative consequences of placement in a track that is above their 'true' performance, is that they have access to shadow education to compensate the gap between their ability and the required performance in the advised track (Elffers, 2018, 2019). We formulate the following hypotheses:

> *H3: Pupils who took the more lenient test, are more likely to end up in a higher academic track in secondary education than pupils who took the traditional test or the stricter test, but this will hold only for pupils from high-SES families.*

3. The Dutch Context

The Dutch Education System

The education system of the Netherlands is a track allocation system, which means that pupils make several transitions within their educational trajectory (Figure 1). The first important transition is in the sixth grade (at age 12) of primary education, when they are allocated to different tracks in secondary education. This transition is based on the teachers' advice and the test advice. Before the test is taken, the primary school teachers give a so-called *initial advice* based on previous performance in school. Then the national test is conducted, and this provides a *test advice.* After the high-stakes test results are available, teachers have the option

to adjust their initial advice and produce a *final advice*. However, the teachers are only allowed to upgrade this advice, not downgrade it.

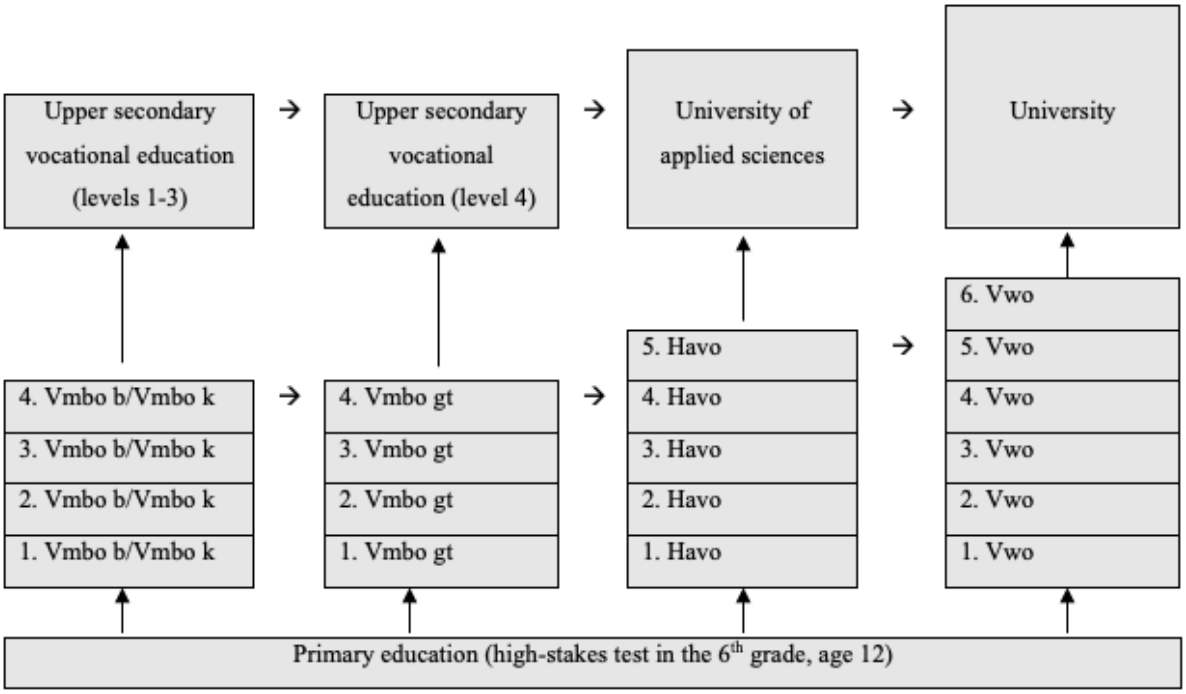

Figure 1. The Dutch educational system and transitions between different tracks

The high-stakes tests

As indicated, the high-stakes tests play an important role in track placement into secondary education. They also have long-term implications because access to tertiary education is based on the diploma obtained from secondary education. Up to 2014/2015, some 90% of the schools in the Netherlands administered the same high-stakes test in the sixth grade, *Eindtoets Basisonderwijs* from supplier Cito, although this was not required. After the new law, the high-stakes test at the end of primary education was mandatory, but primary schools were allowed to choose between three approved high-stakes test suppliers. Although a majority of schools continued using the traditional test, which was renamed in the *Centrale Eindtoets* (CET-test), many schools took this opportunity to switch from the CET-test to one of

the two new tests: the Route 8-test from supplier A-VISION and the *Eindevaluatie Primair Onderwijs* (IEP-test) from supplier Bureau ICE.

There are some differences between the three tests in terms of time and format (for more information, see Appendix 1), but all three aim to provide a track recommendation, based on the test scores in language and math. As mentioned in the introduction, the CET-test has a long tradition, which enabled test developers to calibrate the test annually and develop rules to transform the test scores into a track recommendation. The other two tests were new and, at the time of introduction, not yet calibrated in a high-stakes setting. This means that the cut-off points for different track recommendations could be either too high or too low. This is illustrated in Figures 2a, 2b and 2c that show the distribution in track recommendations for the schools in 2014-2015 and 2015-2016, separately for schools that did not or did switch to a new test supplier.

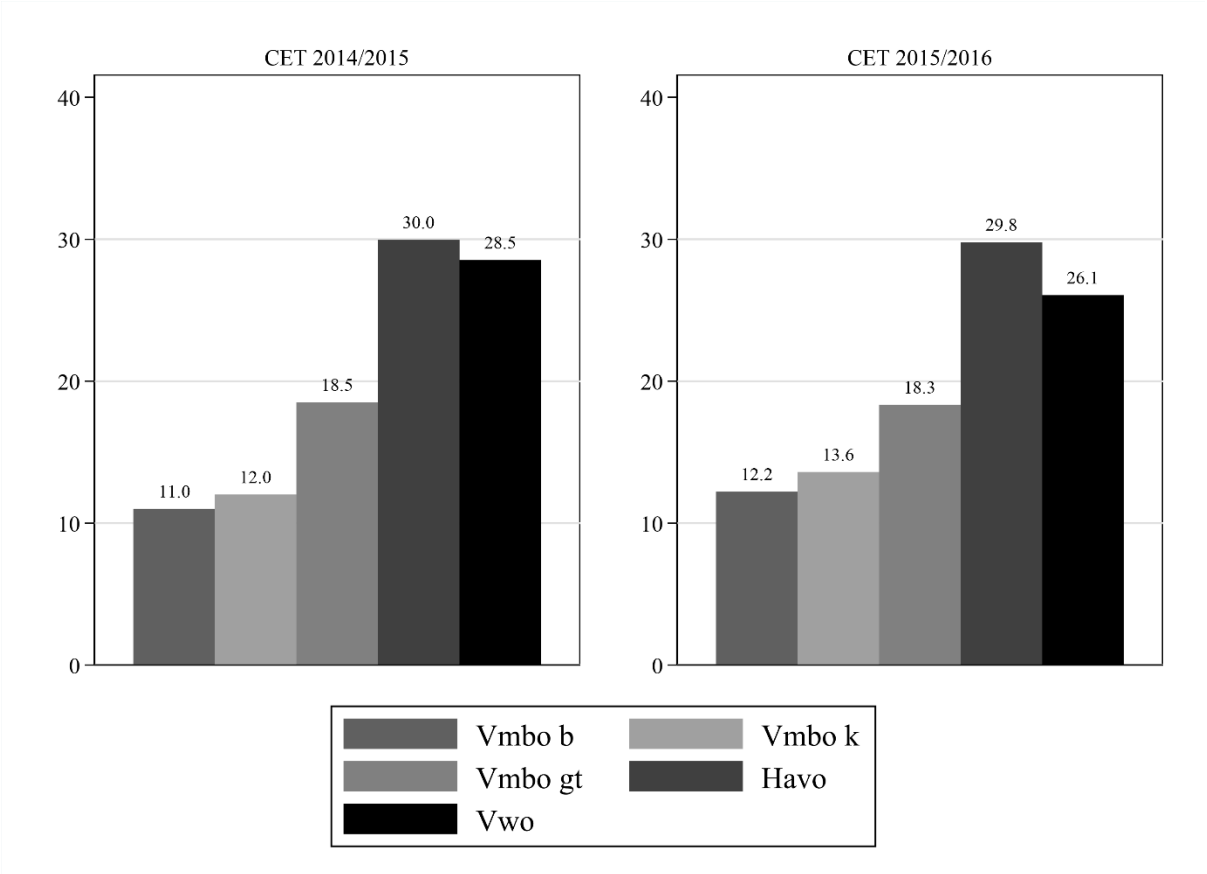

Figure 2a. Distribution of the test advice for schools that did not switch test supplier

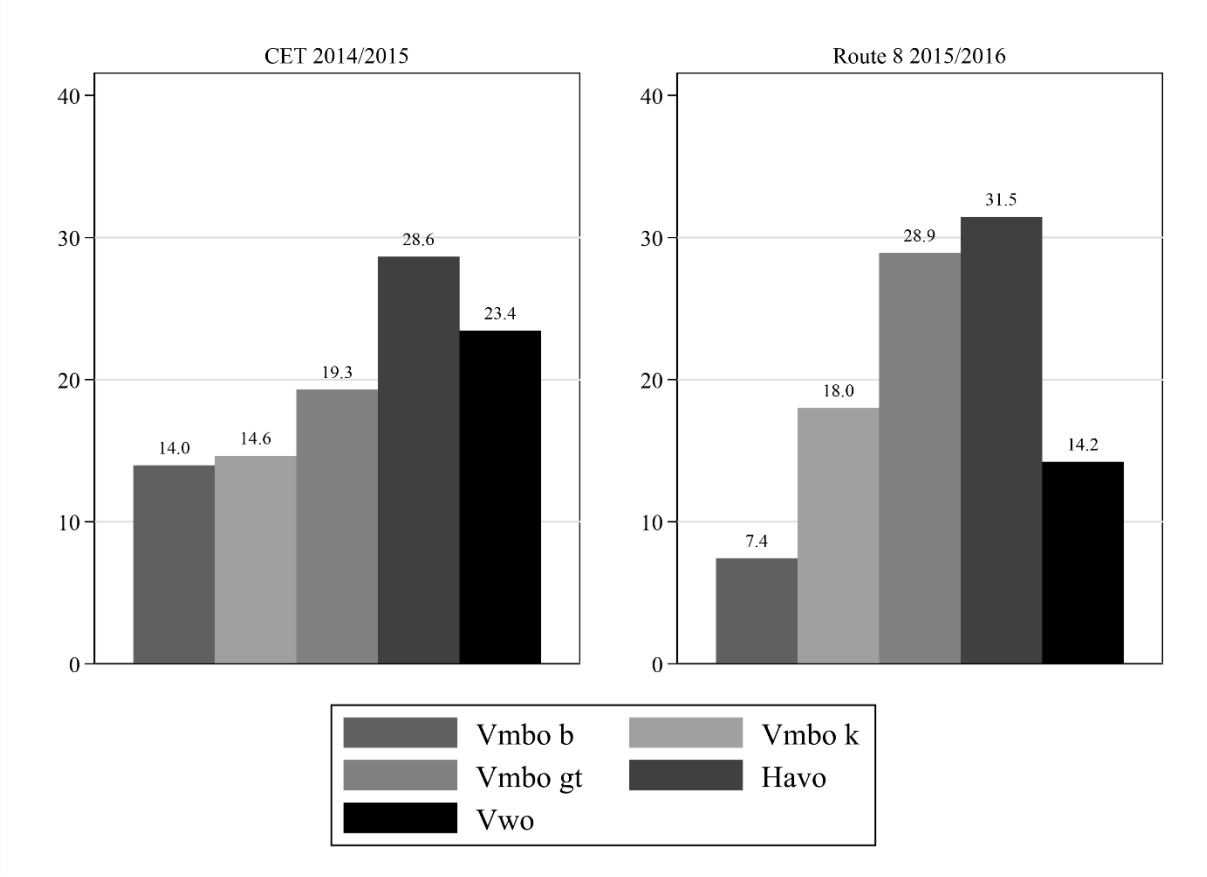

Figure 2b. Distribution of the test advice for schools that switched from the CET-test to the Route 8-test

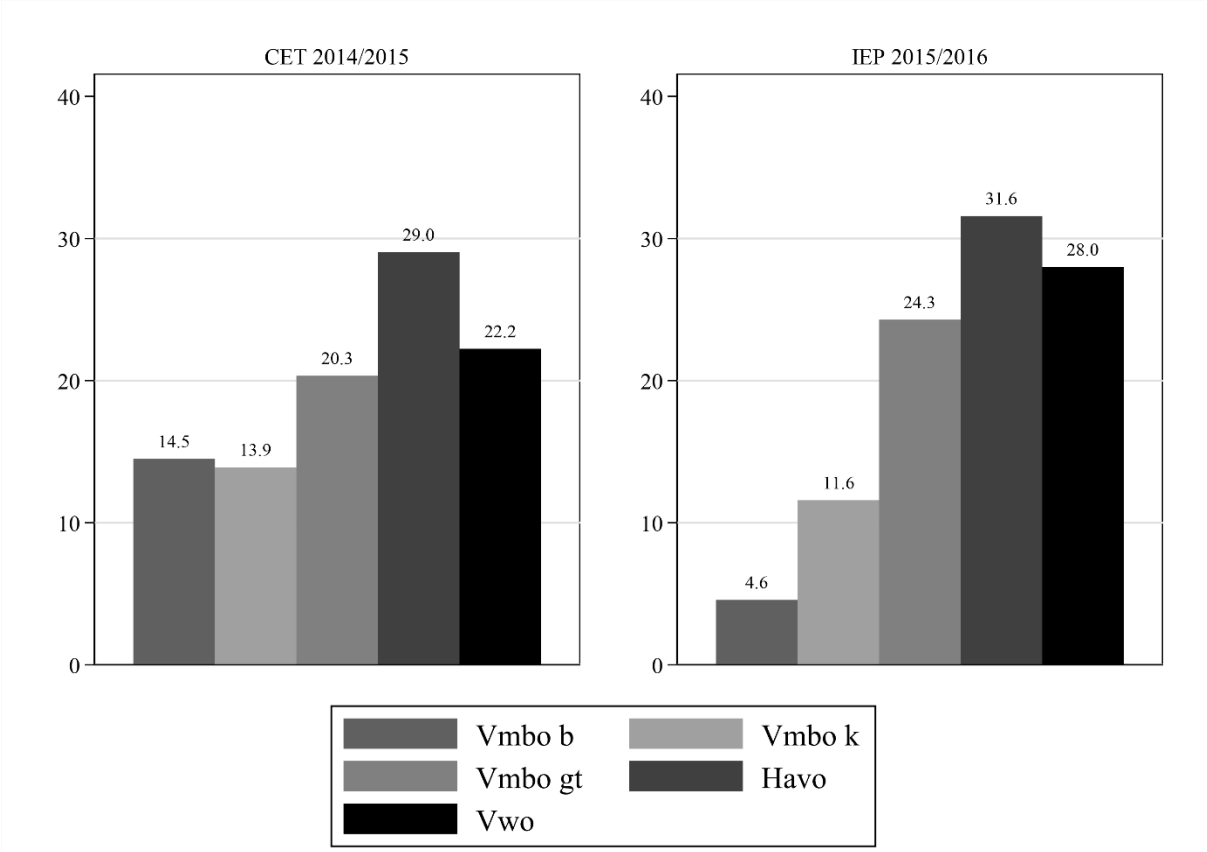

Figure 2c. Distribution of the test advice for schools that switched from the CET-test to the IEP-test

Figure 2a presents the distribution of the test advice of schools that kept using the CET-test. As expected, the distribution of track recommendations between the two different school years is almost the same. There are some differences between the two years, since they are based on other pupils and some fluctuation between years is normal (Bolhaar & Scheer, 2019). Figure 2b shows the distribution for schools that changed from the CET-test to the Route 8-test. The distributions differ strongly between the two years. In 2014/2015, more pupils were advised to go to the lowest track (vmbo b) or highest track (vwo), whereas, in 2015/2016, pupils were more often advised to go to the middle tracks (vmbo k, vmbo gt, and havo). The distribution of the Route 8-test is thus more peaked. Figure 2c shows the distribution for schools that switched to the IEP-test. The distribution changed noticeably between the two years, especially for the highest track. The CET-test of 2014/2015 advised

more pupils to go to the lower tracks (vmbo b and vmbo k) while the IEP-test of 2015/2016 advised more pupils to go to the higher tracks. This confirms that the IEP-test was more lenient compared to the CET-test.

4. Data and Methods

Data

We use data from the NCO dataset (Haelermans et al., 2020). This unique dataset contains register data of all pupils in primary and secondary education in the Netherlands from school year 2008/2009 onwards and includes pupil-, household- and school-characteristics. We selected schools that participated in 2014/2015 in the CET-test and in 2015/2016 in either the Route 8-test, the IEP-test or the CET-test. We deleted cases with missing values for any of the household- and school-characteristics.[2] These selection criteria resulted in an analytical sample of 284,427 pupils, 5,511 schools and 10,946 school year * school combinations.

Dependent Variable

The first dependent variable is whether the results of the high-stakes test led to an *adjusted teacher advice* (dummy 1= yes, 0 = no). Note that the adjustment can only be upward, not downwards.

The second dependent variable is the educational position of the pupil three years after completing primary school based on the educational ladder constructed by Bosker & van der Velden (1989). This scale has been widely used in Dutch educational research and is an easy way to compare different tracks and different grades in one scale (see Figure A1 in

2. Missing data are negligible as we rely on register data.

Appendix 2). The variable is an ordered discrete variable. Table 1 shows the descriptive statistics of the educational position after three years for all pupils in 2014/2015 and 2015/2016.

Table 1. Descriptive statistics of the educational position of pupils after three years

Value	Educational position	N	%
0	Vmbo b (9th grade), equivalent or below	22,207	7.81
1	Vmbo k (9th grade) or equivalent	43,614	15.33
2	Vmbo gt (9th grade) or equivalent	81,901	28.80
3	Havo (9th grade) or equivalent	66,229	23.29
4	Vwo (9th grade), equivalent or above	70,476	24.78
	Total	284,427	100.00

Independent Variables

The *test advice* is based on the score the pupils receive on the different tests. Based on this score the test suppliers determine cut-off points determining the range of scores associated with a certain track recommendation (for an overview, see Table A2 in Appendix 2). The test advice is an ordered discrete variable and has the following values 0 = Vmbo b, 0.5 = Vmbo b/k, 1 = Vmbo k, 1.5 = Vmbo k/gt, 2 = Vmbo gt, 2.5 = Vmbo gt/havo, 3 = Havo, 3.5 = Havo/Vwo, and 4 = Vwo.[3]

The *recalculated test advice* is based on the test advice and has the same values. It is used to estimate the effect of the test advice if the cut-off points in the test would have been the same as in the traditional test. To do this, we use the cut-off points for different track recommendations based on the percentile rank in the distribution of the CET-test in 2014/2015 and use the same percentile ranks as cut-off points for the distribution of the

3. Pupils with advice 'special secondary education' or a level lower than vmbo b (practical education) were omitted from the analyses. They represent only a small percentage of the population and refer mainly to pupils with low cognitive skills or disabilities. Pupils with an advice covering three or more adjacent tracks were also not included in the analyses.

Route 8-test or IEP-test one year later. This is done separately for the schools that switched to the Route 8-test and schools that switched to the IEP-test (results of the distributions available upon request).

Individual Variables

In the analyses, we include the following variables.

Initial teacher advice and *Final teacher advice*, both measured as ordered discrete variables with the same values as the test advice. The correlation between the initial teacher advice and the test advice is $r = 0.8$ (the overlap between initial teacher's advice and test's advice is given in Table A3 of Appendix 2).

The second variable is the *Type of test* with three categories: CET (reference category), Route 8, and IEP.

Other individual variables relate to the background, such as the *Gender* of the pupil (boys as reference category)*; Migration status* (non-migrant as reference category, first-generation migrants and second-generation migrants).

Finally, parental and household characteristics were added. *Father's* and *Mother's employment status*: employed (reference category), receiving benefits, inactive and employment status unknown. *Household structure*: two-parent or one-parent family. Pupils who live without their parents are not included in the analyses. And *Household income*: low household income (lowest 25% as reference category), middle household income (between the 25th and 75th percentiles) and high household income (highest 25%). Unfortunately, the register data do not contain sufficient information about parent's educational attainment or occupational status to determine SES. Instead, we use household income to look at socio-economic differences.

School-Year-Level Variable

The *School year* is either 2014/2015 or 2015/2016. School year 2014/2015 is the reference category. This variable is used in the multilevel analyses to nest pupils in school*year combinations, and school*year combinations into schools.

School-Level Variables

Several primary school indicators were included in the model. First, the school's *Denomination* reflecting the beliefs and vision on which the school operates. In the Netherlands, there are more than 40 different denominations. In the analyses, these are combined into four categories: public schools (reference group), schools based on (educational, pedagogical or societal) philosophies, schools based on religious beliefs and multi-denominational schools. Schools with multiple denominations are often based on pedagogical as well as religious considerations. Furthermore, the index of *Urbanization level* of the area in which the primary school is located, ranging from very low urbanized areas (< 500 addresses/km^2; reference group), low urbanized areas (500 – 1000 addresses/km^2), medium urbanized areas (1000 – 1500 addresses/km^2), strong urbanized areas (1500 – 2500 addresses/km^2) and very strong urbanized areas (>= 2500 addresses/km^2). In addition, the *School size* is added to the model. School sizes are standardized in the analyses.

Table 2 shows the characteristics used in the analyses separately for Route 8-schools, IEP-schools and CET-schools (for total sample see Table A1 of Appendix 2). Generally, the individual-level variables are quite similar across the three types of schools. This means that the student-composition in terms of characteristics such as migration background, household

income, and parental employment status does not differ much between schools that use different test suppliers. However, when we look at school characteristics, we do find some relevant (and statistically significant) differences. For example, Route 8-schools and IEP-schools are located more often in less urbanized areas compared to CET-schools. Looking at the school denomination, we observe that IEP-schools and Route 8-schools are more often public schools, compared to CET-schools.

Table 2. Descriptives for Route 8, IEP and CET, respectively

	CET 2014/2015 - CET 2015/2016			CET 2014/2015 - Route 8 2015/2016			CET 2014/2015 - IEP 2015/2016		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Educational position after 3 years	250,107	2.428	1.230	11,461	2.345	1.236	22,859	2.352	1.231
Test advice	250,107	2.501	1.299	11,461	2.310	1.239	22,859	2.514	1.221
Recalculated test advice	250,107	2.501	1.299	11,461	2.390	1.331	22,859	2.356	1.318
Initial teacher advice	250,107	2.432	1.213	11,461	2.356	1.221	22,859	2.366	1.216
Adjusted teacher advice (No=Ref.)	250,107	0.053	0.223	11,461	0.042	0.200	22,859	0.071	0.256
Final teacher advice	250,107	2.472	1.206	11,461	2.389	1.207	22,859	2.420	1.204
Gender Girls (Boys=ref.)	250,107	0.503	0.500	11,461	0.505	0.500	22,859	0.510	0.500
Migration background (Non-migrant=ref.)	250,107	0.782	0.413	11,461	0.810	0.392	22,859	0.789	0.408
1st generation	250,107	0.020	0.140	11,461	0.021	0.142	22,859	0.021	0.143
2nd generation	250,107	0.198	0.399	11,461	0.169	0.375	22,859	0.191	0.393
Father's employment status (Employed=ref.)	250,107	0.863	0.344	11,461	0.872	0.334	22,859	0.860	0.347
Receives benefit	250,107	0.075	0.263	11,461	0.077	0.266	22,859	0.076	0.264
Inactive	250,107	0.016	0.127	11,461	0.013	0.115	22,859	0.017	0.129
Missing	250,107	0.046	0.210	11,461	0.038	0.191	22,859	0.047	0.212
Mother's employment status (Employed=ref.)	250,107	0.771	0.420	11,461	0.787	0.410	22,859	0.769	0.421
Receives benefit	250,107	0.107	0.309	11,461	0.104	0.305	22,859	0.113	0.317
Inactive	250,107	0.116	0.321	11,461	0.103	0.304	22,859	0.112	0.315
Missing	250,107	0.005	0.073	11,461	0.006	0.078	22,859	0.006	0.079
Household structure 1 adult (2 adults=ref.)	250,107	0.156	0.363	11,461	0.147	0.354	22,859	0.168	0.374
Household income (Low=ref.)	250,107	0.252	0.434	11,461	0.245	0.430	22,859	0.267	0.442
Middle	250,107	0.490	0.500	11,461	0.513	0.500	22,859	0.493	0.500
High	250,107	0.258	0.437	11,461	0.243	0.429	22,859	0.240	0.427
School year 2015/2016 (2014/2015=ref.)	250,107	0.498	0.500	11,461	0.493	0.500	22,859	0.529	0.499
Denomination (Public schools=ref.)	250,107	0.291	0.454	11,461	0.368	0.482	22,859	0.424	0.494
Schools based on philosophies	250,107	0.046	0.209	11,461	0.064	0.244	22,859	0.038	0.190
Schools based on religious beliefs	250,107	0.662	0.473	11,461	0.568	0.495	22,859	0.538	0.499

Multi-denominational	250,107	0.001	0.029	11,461	0.000	0.000	22,859	0.001	0.034
Urbanization (Very low=ref.)	250,107	0.100	0.299	11,461	0.154	0.361	22,859	0.113	0.317
Low	250,107	0.235	0.424	11,461	0.341	0.474	22,859	0.258	0.437
Medium	250,107	0.212	0.408	11,461	0.159	0.366	22,859	0.236	0.424
Strong	250,107	0.280	0.449	11,461	0.222	0.416	22,859	0.249	0.433
Very strong	250,107	0.174	0.379	11,461	0.124	0.330	22,859	0.144	0.352
School size	250,107	303.258	166.598	11,461	302.564	180.442	22,859	280.418	170.670

Methods of Analyses

We use a multilevel design that corrects for the hierarchical clustering of pupils within schools and within school*years (Snijders & Bosker, 2012). We apply a three-level structure in which we nest pupils in school*year combinations and school*year combinations in schools.

Our analyses are estimated in three steps. The first model is a multilevel logistic regression analyses about the test advice and the adjustment in the final teacher advice. The second model is a multilevel linear regression that includes the test advice and the recalculated test advice of the pupils. In the third model, we split the pupils according to their household income (low, middle, high) to see whether the results change across the different socio-economic groups. We estimate the models using the melogit and mixed (linear) package in Stata 15.

5. Results

Relation between the Test Advice and the Upward Adjustment in the Final Teacher Advice

First, we analyze whether the test advice had an impact on the final teachers' advice or rather the upward adjustment in that advice. As indicated above, teachers provided an initial advice before the high-stakes test. The results of the high-stakes test could be used to change that advice, but only in an upward direction. In 2015/2016, approximately 7.0% of our sample had their test advice adjusted: 6.8% of the CET-schools, 5.5% of the Route 8-schools, and 10.4% of the IEP-schools. This result is in line with what we expected, since the IEP-test gave higher track recommendations thus leading to more upward changes. In Table 3, we show the results of a logistic regression analysis whether the teacher advice was adjusted. Models 1-3 present

the results for the original test advice and Models 4-6 for the recalculated test advice. As the effect of changes in the standards of the test directly affect the test advice itself (by offering lower or higher track recommendations), the effect of changing to a new test supplier is best observed in Models 4-6.

Table 3. Multilevel logistic regression analysis of test advice and recalculated test advice on the adjustment of the teacher advice

	Test advice			Recalculated test advice		
	M1	**M2**	**M3**	**M4**	**M5**	**M6**
Test advice	3.161***	3.168***	3.187***			
	(0.027)	(0.026)	(0.027)			
Recalculated test advice				3.055***	3.062***	3.081***
				(0.026)	(0.025)	(0.0126)
Initial teacher advice	-2.968***	-2.973***	-2.999***	-2.932***	-2.938***	-2.964***
	(0.024)	(0.024)	(0.024)	(0.024)	(0.023)	(0.024)
Type of test (CET=ref.)						
Route 8		-0.023	0.029		-0.622***	-0.576**
		(0.185)	(0.184)		(0.184)	(0.183)
IEP		-0.026	0.014		0.766***	0.809***
		(0.116)	(0.115)		(0.115)	(0.115)
School year 2015/2016 (2014/2015=ref.)		1.467***	1.449***		1.451***	1.433***
		(0.050)	(0.050)		(0.050)	(0.050)
Control variables included	No	No	Yes	No	No	Yes
Constant	-6.575***	-7.288***	-7.814***	-6.312***	-7.029***	-7.550***
	(0.060)	(0.068)	(0.123)	(0.058)	(0.066)	(0.121)
School variance	1.461***	2.109***	1.878***	1.432***	2.080***	1.858***
	(0.137)	(0.127)	(0.121)	(0.138)	(0.125)	(0.120)
School year variance	3.885***	2.606***	2.615***	3.942***	2.549***	2.557***
	(0.168)	(0.120)	(0.121)	(0.171)	(0.118)	(0.118)
N pupils	284,427	284,427	284,427	284,427	284,427	284,427
N Schools	5,511	5,511	5,511	5,511	5,511	5,511
N Schools*School year	10,946	10,946	10,946	10,946	10,946	10,946

* p<0.05, ** p<0.01, *** p<0.001; Standard errors in parentheses; Models 3 and 6 are controlled for gender, migration background, father's and mother's employment status, household structure, household income, denomination and urbanization of school and school size; Results for the full model available upon request.

Across all models, we find that higher scores on the (recalculated) test advice are correlated with a greater likelihood that teacher adjust their advice. This means that pupils with a higher (recalculated) test advice are more likely to get an adjusted test advice. Additionally, we see that a lower initial teacher advice is correlated with lower chances of getting an adjusted teacher advice.

As indicated, the effect of the different types of tests is best observed in models 5 and 6. Controlled for the recalculated test advice, pupils who took the IEP-test have a much higher chance to receive an upward adjusted advice than the control group who took the CET-test (0.77 in the model 5 and 0.81 in the model 6). The opposite holds for pupils who took the Route 8-test: They have a lower chance to receive an upward advice (-0.62 and -0.58 respectively). This means that *H1* is confirmed: The initial teacher advice of pupils who took the more lenient test (IEP), is more often adjusted in an upward direction.

Table 4. Multilevel logistic regression analysis of test advice and recalculated test advice on the adjustment of the teacher advice for different income groups

	M6a **Low income**	**M6b** **Middle income**	**M6c** **High income**
Recalculated test advice	3.075***	2.887***	3.485***
	(0.0250)	(0.036)	(0.070)
Initial teacher advice	-2.869***	-2.801***	-3.473***
	(0.048)	(0.033)	(0.063)
Type of test (CET=ref.)			
Route 8	-0.380	-0.635**	-0.894**
	(0.247)	(0.197)	(0.277)
IEP	0.853***	0.749***	0.916***
	(0.147)	(0.121)	(0.163)
School year 2015/2016 (2014/2015=ref.)	1.448***	1.363***	1.251***
	(0.073)	(0.055)	(0.077)
Control variables included	Yes	Yes	Yes
Constant	-7.646***	-6.982***	-7.076***
	(0.183)	(0.132)	(0.203)
School variance	1.583***	1.437***	1.657***
	(0.183)	(0.126)	(0.209)
School year variance	2.651***	2.151***	2.371***
	(0.213)	(0.136)	(0.239)
N pupils	71,875	139,814	72,738
N Schools	5,441	5,505	5,227
N Schools*School year	10,391	108,44	9,692

* p<0.05, ** p<0.01, *** p<0.001; Standard errors in parentheses; Models are controlled for gender, migration background, father's and mother's employment status, household structure, denomination and urbanization of school and school size; Results for the full model available upon request.

Does this conclusion also hold for the different income groups? In Table 4, we present the results of our preferred model with the recalculated test advice of Table 3 (Model 6), separately for pupils from low, middle, and high household incomes. For pupils who took the IEP-test, we find no differences across the three income levels: All groups profit equally from having taken the IEP-test. This means that *H2* is refuted. For the Route 8-test, we note that the negative effect is only observed for pupils from the middle- and high-income families. This makes sense as these pupils are more often eligible for the highest academic track (VWO) for

which the Route 8-test less often advices (see Figure 2b). The opposite holds for the pupils from low-income families: They also suffer from the under-advising for the highest academic track, but this is 'compensated' by similar under-advising for the lowest track (VMBO b, see Figure 2b). The overall effect is therefore not significant.

Relation between the Test Advice and the Educational Position after Three Years

Table 5 presents the multilevel regression estimates of the relation between the test advice and the educational position after three years. Again, we show separate models for the test advice and the recalculated test advice. Across all models, we find that a higher score on the (recalculated) test advice as well as on the initial teacher advice correlates strongly with a higher educational position after three years.

Regarding the test advice, we see in Model2 2 and 3 that the pupils who took the Route 8-test do not significantly differ from those who took the CET-test. Since we also added a school year variable, the comparison with the CET-test is in fact a comparison with the pupils who took the CET-test in 2014/2015. Pupils who took a Route 8-test do not have a higher educational position in secondary education after three years, compared to pupils in the previous cohort of the same schools who took the CET-test. When we look at the recalculated test advice models, we still do not observe a significant effect of having taken the Route 8-test (vs. CET-test) on the educational position three years later.

For the IEP-test however, we find a significant negative association: Pupils who took the IEP-test are less likely to end up with a high position in secondary education three years later, compared to the previous cohort of pupils in the same schools who took the CET-test (Models 2 and 3). The negative effect of -0.091 (Model 3) means that pupils who took this test dropped by one 10th of a school level once they reach the third year in secondary

education (grade 9). However, when looking at the recalculated test advice models, we can see that this is entirely due to the IEP-test track recommendations being systematically too high. If we control for this by using the recalculated advice, the negative effect is barely significant (Model 5) or not significant (Model 6). This means that there is no positive long-term effect of having taken the more lenient test.

Table 5. Multilevel linear regression analysis of test advice and recalculated test advice on the educational position after three years

	Test advice			Recalculated test advice		
	M1	**M2**	**M3**	**M4**	**M5**	**M6**
Test advice	0.274***	0.275***	0.269***			
	(0.002)	(0.002)	(0.002)			
Recalculated test advice				0.273***	0.274***	0.267***
				(0.002)	(0.002)	(0.002)
Initial teacher advice	0.628***	0.626***	0.617***	0.626***	0.625***	0.616***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Type of test (CET=ref.)						
Route 8		0.021	0.022		-0.022	-0.020
		(0.012)	(0.011)		(0.012)	(0.011)
IEP		-0.101***	-0.091***		-0.018*	-0.010
		(0.008)	(0.008)		(0.008)	(0.008)
School year 2015/2016 (2014/2015=ref.)		0.016***	0.014***		0.016***	0.014***
		(0.003)	(0.003)		(0.003)	(0.003)
Control variables included	No	No	Yes	No	No	Yes
Constant	0.194***	0.190***	0.074***	0.203***	0.196***	0.080***
	(0.003)	(0.003)	(0.008)	(0.003)	(0.003)	(0.008)
School variance	0.018***	0.018***	0.013***	0.018***	0.018***	0.013***
	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)
School year variance	0.005***	0.005***	0.005***	0.005***	0.005***	0.005***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Residual variance	0.319***	0.319***	0.306***	0.319***	0.319***	0.306***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N pupils	284,427	284,427	284,427	284,427	284,427	284,427
N Schools	5,511	5,511	5,511	5,511	5,511	5,511
N Schools*School year	10,946	10,946	10,946	10,946	10,946	10,946

* p<0.05, ** p<0.01, *** p<0.001; Standard errors in parentheses; Models 3 and 6 are controlled for gender, migration background, father's and mother's employment status, household structure, household income, denomination and urbanization of school and school size; Results for the full model available upon request

We now test whether these conclusions also hold for the different subgroups. In Table 6, we present the results of our preferred model with the recalculated test advice of Table 5 (Model 6), separately for pupils with low, middle, and high household incomes. Basically, we find that the type of test has no effect at all for any income group, so we find no support for *H3*: There

are no long-term benefits from having taken the more lenient test for the pupils from the high-income groups.

Table 6. Multilevel linear regression analysis of recalculated test advice on educational position after three years for different income groups

	M6a Low income	**M6b Middle income**	**M6c High income**
Recalculated test advice	0.273***	0.254***	0.267***
	(0.003)	(0.002)	(0.003)
Initial teacher advice	0.596***	0.633***	0.623***
	(0.003)	(0.002)	(0.003)
Type of test (CET=ref.)			
Route 8	-0.020	-0.014	-0.015
	(0.020)	(0.014)	(0.017)
IEP	-0.017	0.002	0.004
	(0.013)	(0.010)	(0.012)
School year 2015/2016 (2014/2015=ref.)	0.024***	0.012***	0.002
	(0.005)	(0.004)	(0.004)
Control variables included	Yes	Yes	Yes
Constant	0.093***	0.151***	0.253***
	(0.012)	(0.009)	(0.011)
School variance	0.014***	0.013***	0.011***
	(0.001)	(0.001)	(0.001)
School year variance	0.007***	0.005***	0.003***
	(0.001)	(0.001)	(0.001)
Residual variance	0.361***	0.304***	0.253***
	(0.002)	(0.001)	(0.001)
N pupils	71,875	139,814	72,738
N Schools	5,441	5,505	5,227
N Schools*School year	10,391	10,844	9,692

* p<0.05, ** p<0.01, *** p<0.001; Standard errors in parentheses; Models are controlled for gender, migration background, father's and mother's employment status, household structure, denomination and urbanization of school and school size; Results for the full model available upon request

Robustness Checks

We ran several robustness checks to scrutinize our findings. First, we split our sample into two groups based on the test score and compare the pupils with above and below median

scores in 2014/2015 with those in 2015/2016 in the same schools (see Table A4 in Appendix 2). We ran our preferred model for these two groups separately. Interestingly, we observe a negative effect for both IEP and Route 8-test for pupils with above median test scores. For pupils who took the Route 8-test, we find a negative effect of -0.038. This reflects the fact that for the above median pupils who took the Route 8, the stricter test more often denied them access to an appropriate track. Although the effect is not very substantial (about 1/26th of a school level difference), it is significant at the p=0.01 level. We also find a similar negative effect for pupils who took the more lenient IEP-test. This seems to suggest that for the above median pupils, having taken the lenient test might even have harmed their subsequent career, despite the initial positive effect on the teacher's advice. In both cases however, the effect is not very substantial.

Second, we split our sample into pupils who received a vmbo b, vmbo k, vmbo gt, havo or vwo test track recommendation, to observe whether the type of test had any effect for a specific range of test scores (corresponding to a certain test advice). Here we only use single recalculated test advices. The results are presented in Table A5 in the Appendix 2. Except for a HAVO-advice for Route 8, we find no long-term effects of the type of test for specific track recommendations on the educational position three years later.

6. Conclusion and Discussion

High-stakes tests are often used to allocate pupils to prestigious tracks or schools in secondary education or students to prestigious programs or colleges in tertiary education. It is not clear what would happen if the standards for such tests would be systematically lowered for one group or raised for another group. Would the former group profit from this? And if so, does

this last? And what about the detrimental effects for the group for whom the standards were raised?

Theoretically one could expect positive outcomes for the group for whom the bar was lowered. Being signaled in the test as a 'high performer' (even if this not entirely true) might raise expectations and result in a self-fulfilling prophecy. And in the opposite case, being signaled as a 'low performer' on the test and being denied access to a track that is too low might result in motivational problems and dropout. We might also expect that these effects might differ between socio-economic groups. Parents and pupils from a high-SES background might find it easier to use the opportunities provided by lower standards or circumvent the obstacles of higher standards.

Although these questions are highly relevant for policymakers to design admission policies, it is hard to get experimental evidence on the long-term consequences. Running an experiment with lowering the standards for some pupils but not for others would be considered unethical because of the potential huge implications for pupils' careers. In this paper we use a unique natural experiment to assess how the standards of a high-stakes test at the end of primary education affect a pupil's performance in secondary education. Traditionally, some 90% of the schools in the Netherlands used the same high-stakes test, namely, the CET-test. We employ a change in the law allowing two new suppliers of high-stakes tests to enter the market: Route 8 and IEP. All three tests convert the test score in a so-called track recommendation. This track recommendation (referred to as test advice) plays an important role in the primary school teacher's advice and the initial track placement in secondary education. We use the fact that, in the year of introduction, the new tests were not yet properly calibrated as they were developed and calibrated on a small sample of pupils in a low-stakes setting. This implied that the cut-off points for the different track

recommendations were too high in one case and too low in the other case. The IEP-test was systematically converting test scores in track recommendations that were too high, and the other new test, the Route 8-test, was giving track recommendations that were too low for the high-achieving group. It is important to note that schools had no prior knowledge on these characteristics of the test and therefore this cannot have played a role in the decision of the school to switch to one of the new test suppliers.

We use high-quality register data from the NCO, covering some 285,000 pupils from over 5,500 primary schools. We use a within-schools-across-cohorts multilevel design to model the short- and long-term outcomes. Does the standard of the test affect the final teacher's advice, and does it affect educational position three years later? And are these effects heterogeneous across pupils from different income groups?

We find that pupils who took the IEP-test initially profited from this by receiving a higher final teacher advice. However, after three years in secondary education, this did not result in a higher track. Instead, these pupils ended up in the same educational position as a control group taking the traditional test. The same holds for the Route 8-test, although the expectations for this test were different since the teachers could only upgrade and not downgrade their initial advice. This means that pupils who took the Route 8-test did not suffer from it, even though this test more often advised pupils to lower tracks than would have been the case if the pupils had taken the traditional test. However, we did find some indications that the pupils with above median test scores, did experience some negative effect from having taken the lenient (IEP) as well as the stricter test (Route 8). However, these effects sizes are quite small. We also conducted these analyses separately by social group but found no differences between pupils from low-, middle- and high-income groups.

Implications

What are the policy implications? First, we can interpret the results as showing that the Dutch educational system is able to correct for mistakes in the allocation process. Even though the IEP-test clearly led a large proportion of pupils to get a final teacher advice that is too high, this was compensated for or corrected in the first years of secondary education. In that sense, there is enough flexibility in the system to compensate for weak links in the allocation process chain. This is in line with a similar conclusion by Dustmann, Puhani & Schonberg (2017) for Germany, a country also known for its early stratification. They examine the long-term outcomes of misallocation in the transition from primary to secondary education in Germany and find no effects on wages, employment or occupation choice at later ages: *"These findings emphasise a core aspect of the basis on which tracking systems should be assessed: the built-in possibilities for correcting earlier allocations at a stage when more information is revealed about a student's true potential." (*Dustmann et al., 2017: 1348). The results suggest that the Dutch education also has these built-in flexibilities, that allow earlier errors in the track placement to be corrected. This flexibility is an important but also overlooked feature that characterizes education systems (Wessling & Van der Velden, 2021).

It is important to note that this is only possible, because schools in the Netherlands have the option of postponing the actual track decision by offering so-called bridging classes. In such schools, the actual tracking can be postponed to age 13 or even age 14. This means that any mistakes, either in test advice or final teacher advice, can be corrected in the first two years. In that sense, it is worrisome that in the past few years more and more schools are switching to homogeneous tracks as of the first year in secondary education (Inspectorate of Education, 2018) which has negative implications for providing opportunities to pupils to

reach their full potential, as well as a successful school career (Bles, Van der Velden, & Ariës, 2020).

The results also indicate that there is no easy solution to increase the enrollment of disadvantaged groups in secondary education. Affirmative action, such as giving low-SES children higher track recommendations will not automatically result in better educational careers if this track recommendation is not accompanied by extra support in secondary education (Baker & Johnston, 2010; Bodvin, Verschueren, De Haene, & Struyf, 2018).

A final interesting conclusion is that the initial teacher's advice is less sensitive to systematic errors than the test advices are. Looking at the results, we can conclude that the new IEP-test did lead to a change (upward adjustment) in teachers' advice, but in this case the initial advice was better than the final advice.

Possible limitations

Ideally it would have been best to design an RCT to assess the effects of changes in the standards of a test on subsequent allocation. As indicated, such a design is unlikely to be approved by an ethical committee because of the potential negative consequences for pupils involved. In the absence of an RCT, we think that this natural experiment comes very close. The schools and the teachers had no prior knowledge that the standards of these tests were different. Thus, there is no reason to believe that the standards of the test played a role in the decision to switch to a new supplier. This is in line with the fact that some schools switched to a test that was more lenient while other schools switched to a test that was stricter.

Still, it could be argued that schools self-select into new test suppliers. We address this by using a within-schools-across-cohorts multilevel design. This means that we compare cohorts within the same schools that got a different treatment. It is very unlikely that the compositions of pupil cohorts within a school changes over time.

Another issue that often plagues experimental research is selective panel mortality. In this case however, we use high quality register data that includes all schools.

References

Baker, M., & Johnston, P. (2010). The impact of socioeconomic status on high stakes testing reexamined. *Journal of Instructional Psychology, 37*(3), 193-199.

Bles, P., Van der Velden, R., & Ariës, R. (2020). Is there an opportunity-performance trade-off in secondary education? Maastricht: *ROA Research Memorandum 9*.

Bodvin, K., Verschueren, K., De Haene, L., & Struyf, E. (2018). Social inequality in education and the use of extramural support services: access and parental experiences in disadvantaged families. *European Journal of Psychology of Education, 33*(2), 215-233.

Bolhaar, J., & Scheer, B. (2019). Verschil in leerresultaten basisscholen. *CPB Notitie*. Retrieved from https://www.cpb.nl/verschillen-in-leerresultaten-tussen-basisscholen

Boudon, R. (1974). *Education, opportunity and social inequality*. New York: Wiley.

Bosker, R. J., & Van der Velden, R. K. W. (1989). Schooleffecten en rendementen. In: J. van Damme & J. Dronkers (Eds.), *Jongeren in school en beroep* (pp. 25-40). Amsterdam: Swets, Zeitlinger.

Breen, R., & Goldthorpe, J. (1997). Explaining educational differentials: Towards a formal rational action theory. *Rationality and Society, 9*, 275–305.

Breen, R., Van De Werfhorst, H. G., & Jæger, M. M. (2014). Deciding under doubt: A theory of risk aversion, time discounting preferences, and educational decision-making. *European Sociological Review, 30*(2), 258-270.

Dumont, H., Klinge, D., & Maaz, K. (2019). The Many (Subtle) Ways Parents Game the System: Mixed-Method Evidence on the Transition into Secondary-School Tracks in Germany. *Sociology of education, 92*(2), 199-228.

Dustmann, C., Puhani, P., & Schonberg, U. (2017). The long-term effects of early track choice. *The Economic Journal, 127*(603), 1348-1380.

Elffers, L. (2018). *De bijlesgeneratie: Opkomst van de onderwijscompetitie*. Amsterdam: Amsterdam University Press.

Elffers, L. (2019). Het kopen van kansen: De inzet van schaduwonderwijs in de onderwijscompetitie. In: H. Van de Werfhorst & E. Van Hest (Eds.), *Gelijke kansen in de stad*. Amsterdam: Amsterdam University Press.

Elsner, B., & Isphording, I. (2017). A Big Fish in a Small Pond: Ability Rank and Human Capital Investment. *Journal of Labor economics, 35*(3), 787-828.

Forster, A., & Van de Werfhorst, H. (2020). Navigating Institutions: Parents' Knowledge of the Educational System and Students' Success in Education. *European Sociological Review, 36*(1), 48-67.

Haelermans, C., Huijgen, T., Jacobs, M., Levels, M., Van der Velden, R., Van Vugt, L., & Van Wetten, S. (2020). Using data to advance educational research, policy and practice: Design, content and research potential of the Netherlands Cohort Study on Education. *European Sociological Review, 36*(4), 643-622.

Hardy, L. (2003). Overburdened, overwhelmed. *American School Board Journal, 190*(4), 18-23.

Inspectorate of Education. (2018). *De Staat van het Onderwijs 2018* Utrecht: Inspectie van het Onderwijs

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.

Taylor, M. C. (1979). Race, sex and the expression of self-fulfilling prophecies in a laboratory teaching situation. *Journal of personality and social psychology, 37*(6), 871-912.

Timmermans, A. C., de Boer, H., Amsing, H. T. A., & van der Werf, M. P. C. (2018). Track recommendation bias: Gender, migration background and SES bias over a 20-year period in the Dutch context. British Educational Research Journal, 44(5), 847-874.

Vaisey, S. (2006). Education and its discontents: Overqualification in America 1972-2002. *Social Forces, 85*(2), 835-864.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge: Harvard University Press.

Wessling, K. & Van der Velden, R. (2021). Flexibility in educational systems - Concept, indicators, and directions for future research. Maastricht: *ROA Research Memorandum No. 002*

Wilson, R. C., Shenhav, A., Straccia, M., & Cohen, J. D. (2019). The Eighty Five Percent Rule for optimal learning. *Nature communications, 10*(1), 1-9.

Appendix 1: Background information on the tests

CET-test

The CET-test entails three mornings/afternoons of two or three hours in which the pupils receive questions related to math and language. Almost all the pupils take the CET-test on paper. The CET-test is not adaptive, but during the period we analyze, there were two versions of the test, the so-called N-test and B-test. The N-test is for pupils the teacher expects to follow theoretical pre-vocational tracks or higher (vmbo gt, in Dutch or higher). The B-test is for pupils the teacher expects would be most suited to the practical pre-vocational tracks (vmbo b or vmbo k, in Dutch). This decision is not meant to be a form of pre-selection but, rather, to make sure that all pupils are assessed as precisely as possible with a test that is neither too easy nor too difficult. Since both tests partly overlap (i.e., a quarter of the test items are the same), the results of the B-test can be transformed into track recommendations corresponding to the N-test so that all the track recommendations are comparable.

Route 8-test

The Route 8-test differs from the CET-test because it is an adaptive computer test, meaning that it responds to the quality of pupils' answers. Pupils are given different difficulty-level dependent questions based on their performance, so that the questions match their ability. Furthermore, the test is a much shorter, encompassing only about two to three hours, whereas the CET-test takes about six to nine hours.

IEP-test

The IEP-test takes about four hours, spread over two mornings, and is therefore a shorter test than the CET-test. Furthermore, just like the CET-test, the IEP-test is administered on paper.

Appendix 2: Background tables and figures

Table A1. Descriptives of the total sample

	N	Mean	SD	Min	Max
Educational position after 3 years	284,427	2.419	1.231	0	4
Test advice	284,427	2.495	1.291	0	4
Recalculated test advice	284,427	2.485	1.303	0	4
Initial teacher advice	284,427	2.424	1.214	0	4
Adjusted teacher advice (No=Ref.)	284,427	0.054	0.225	0	1
Final teacher advice	284,427	2.465	1.206	0	4
Type of test (CET=ref.)	284,427	0.938	0.242	0	1
Route 8	284,427	0.020	0.140	0	1
IEP	284,427	0.043	0.202	0	1
Gender Girls (Boys=ref.)	284,427	0.504	0.500	0	1
Migration background (Non-migrant=ref.)	284,427	0.784	0.412	0	1
1st generation	284,427	0.020	0.140	0	1
2nd generation	284,427	0.196	0.397	0	1
Father's employment status (Employed=ref.)	284,427	0.863	0.344	0	1
Receives benefit	284,427	0.075	0.263	0	1
Inactive	284,427	0.016	0.126	0	1
Missing	284,427	0.046	0.209	0	1
Mother's employment status (Employed=ref.)	284,427	0.772	0.420	0	1
Receives benefit	284,427	0.107	0.309	0	1
Inactive	284,427	0.115	0.320	0	1
Missing	284,427	0.006	0.074	0	1
Household structure 1 adult (2 adults=ref.)	284,427	0.157	0.364	0	1
Household income (Low=ref.)	284,427	0.253	0.435	0	1
Middle	284,427	0.492	0.500	0	1
High	284,427	0.256	0.436	0	1
School year 2015/2016 (2014/2015=ref.)	284,427	0.500	0.500	0	1
Denomination (Public schools=ref.)	284,427	0.305	0.460	0	1
Schools based on philosophies	284,427	0.046	0.209	0	1
Schools based on religious beliefs	284,427	0.648	0.477	0	1
Multi-denominational	284,427	0.001	0.029	0	1
Urbanization (Very low=ref.)	284,427	0.103	0.304	0	1
Low	284,427	0.241	0.428	0	1
Medium	284,427	0.211	0.408	0	1
Strong	284,427	0.275	0.447	0	1
Very strong	284,427	0.169	0.375	0	1
School size	284,427	301.394	167.622	12	1,283

Table A2. Overview of categorisation of test scores into test advices

CET-test		**Route 8-test**		**IEP-test**	
2014/2015					
501–518	vmbo b				
519–525	vmbo b/k				
526–528	vmbo k				
529–532	vmbo gt				
533–536	vmbo gt, havo				
537–539	havo				
540–544	havo / vwo				
545–550	vwo				
2015/2016		*2015/2016*		*2015/2016*	
501 – 518	vmbo b	141 - 168	vmbo b	50-61	vmbo b/k
519 – 525	vmbo b/k	169 - 190	vmbo k	62-70	vmbo k/tl(gt)
526 – 528	vmbo k	191 - 210	vmbo gt	71-76	vmbo gt
529 – 532	vmbo gt	211 - 234	havo	77-81	vmbo gt/havo
533 – 536	vmbo gt, havo	>234	vwo	82-86	havo
537 – 539	havo			87-92	havo/vwo
540 – 544	havo / vwo			93-100	vwo
545 – 550	vwo				

Table A3. Crosstabulation of initial teacher advice and test advice in percentages

		Initial teacher advice									
		Vmbo bb	**Vmbo bb/kb**	**Vmbo kb**	**Vmbo kb/gt**	**Vmbo gt**	**Vmbo gt/havo**	**Havo**	**Havo/vwo**	**Vwo**	**Total**
Test advice	**Vmbo bb**	58.61	11.65	21.67	1.74	5.78	0.31	0.16	0.03	0.04	100
	Vmbo bb/kb	23.01	9.07	35.75	4.60	24.21	1.90	1.25	0.16	0.06	100
	Vmbo kb	8.18	5.40	30.58	5.49	40.30	5.01	4.58	0.28	0.18	100
	Vmbo kb/gt	15.67	8.33	34.17	6.57	28.59	3.67	2.75	0.15	0.08	100
	Vmbo gt	3.16	2.70	20.35	4.52	46.78	8.45	12.13	1.33	0.59	100
	Vmbo gt/havo	0.92	1.01	9.62	2.69	41.19	12.00	26.94	3.68	1.95	100
	Havo	0.27	0.35	3.31	1.21	25.30	11.01	41.83	8.74	8.00	100
	Havo/vwo	0.06	0.10	0.76	0.34	9.58	5.85	43.45	15.23	24.63	100
	Vwo	0.01	0.01	0.08	0.04	1.15	0.91	15.32	10.74	71.74	100
	Total	7.36	2.64	11.79	2.14	22.21	5.81	21.49	6.61	19.93	100

Value	Educational position after three years				
4	**Vwo 9th grade**	Havo 10th grade			
3	Vwo 8th grade	**Havo 9th grade**	Vmbo gt 10th grade		
2	Vwo 7th grade	Havo 8th grade	**Vmbo gt 9th grade**	Vmbo k 10th grade	
1		Havo 7th grade	Vmbo gt 8th grade	**Vmbo k 9th grade**	Vmbo b 10th grade
0			Vmbo gt 7th grade	Vmbo k 8th grade	**Vmbo b 9th grade**

Figure A1. Educational positions of pupils after three years on the educational ladder

Table A4. Multilevel linear regression analyses of test advice on the educational position after 3 years based on pupils with a percentile score below median vs. pupils with a median score or above

	M6a Below median	**M6b Median or above**
Recalculated test advice	0.226***	0.298***
	(0.002)	(0.003)
Initial teacher advice	0.595***	0.627***
	(0.002)	(0.003)
Type of test (CET=ref.)		
Route 8	-0.019	-0.038**
	(0.015)	(0.014)
IEP	0.009	-0.035***
	(0.010)	(0.010)
School year 2015/2016 (2014/2015=ref.)	0.026***	-0.005
	(0.004)	(0.004)
Control variables included	Yes	Yes
Constant	0.146***	-0.037**
	(0.010)	(0.012)
School variance	0.015***	0.015***
	(0.001)	(0.001)
School year variance	0.005***	0.006***
	(0.001)	(0.001)
Residual variance	0.318***	0.287***
	(0.001)	(0.001)
N pupils	139,821	144,606
N Schools	5,488	5,482
N Schools*School year	10,149	10,744

* p<0.05, ** p<0.01, ** p<0.001; Standard errors in parentheses; All models are controlled for gender, immigration background, father's and mother's employment status, household structure, household income, denomination and urbanization of school and school size

Table A5. Multilevel linear regression analyses of test advice on the educational position after 3 years by different treatment groups based on single recalculated test advice

	M6 vmbo b	**M6 vmbo k**	**M6 vmbo gt**	**M6 havo**	**M6 vwo**
Initial teacher advice	0.597***	0.601***	0.573***	0.618***	0.588***
	(0.006)	(0.005)	(0.004)	(0.005)	(0.004)
Type of test (CET=ref.)					
Route 8	0.046	-0.010	-0.048	-0.060*	-0.027
	(0.026)	(0.031)	(0.026)	(0.028)	(0.016)
IEP	0.011	0.021	0.020	-0.033	-0.007
	(0.019)	(0.020)	(0.019)	(0.019)	(0.012)
School year 2015/2016 (2014/2015=ref.)	-0.002	0.018*	0.027***	0.032***	-0.001
	(0.008)	(0.008)	(0.007)	(0.008)	(0.004)
Control variables included	Yes	Yes	Yes	Yes	Yes
Constant	0.152***	0.420***	0.603***	0.752***	1.391***
	(0.018)	(0.020)	(0.018)	(0.022)	(0.018)
School variance	0.009***	0.014***	0.014***	0.021***	0.007***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
School year variance	0.006***	0.006***	0.011***	0.007***	0.005***
	(0.002)	(0.003)	(0.002)	(0.002)	(0.001)
Residual variance	0.237***	0.312***	0.319***	0.358***	0.182***
	(0.003)	(0.004)	(0.003)	(0.003)	(0.001)
N pupils	18,755	21,307	33,384	31,266	56,465
N Schools	4,568	4,992	5,296	5,268	5,285
N Schools*School year	7,127	8,323	9,517	9,351	9,696

* p<0.05, ** p<0.01, ** p<0.001; Standard errors in parentheses; All models are controlled for gender, immigration background, father's and mother's employment status, household structure, household income, denomination and urbanization of school and school size